LAVA #04

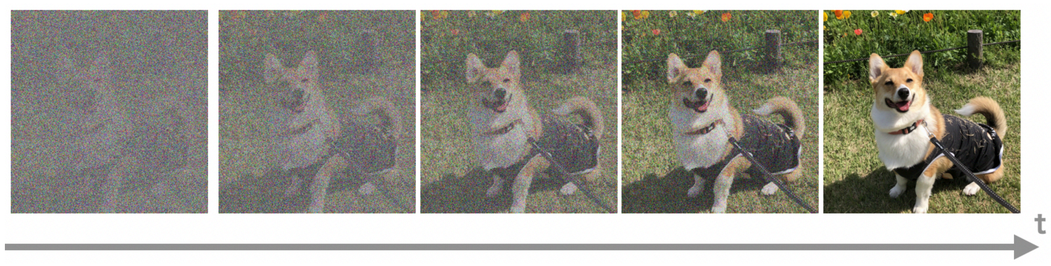
Exploring Cross-Attention Maps in Multi-modal Diffusion Transformers for Training-Free Semantic Segmentation

Rento Yamaguchi, Keiji Yanai

The University of Electro-Communications

1. Background

Diffusion models can generate high-quality images from noise, with Cross Attention Maps capturing object positions from prompt tokens.



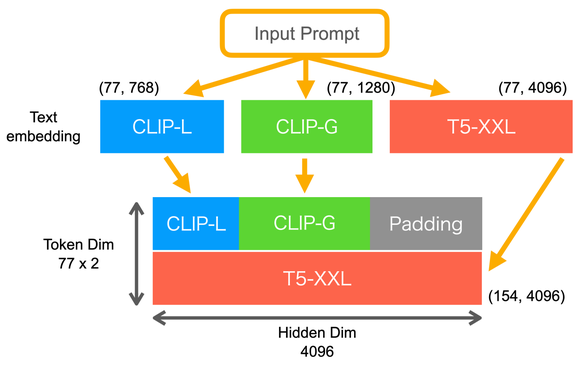
Generation Process of the Diffusion Model

Token and Image Region Relationships in Cross Attention Maps

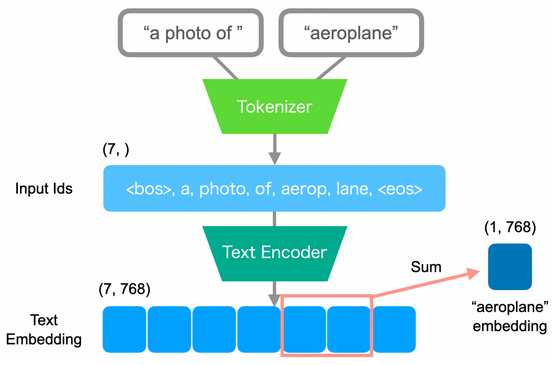
We investigated the object grouping capability in the cross attention maps of recent high-quality image generators based on Diffusion Transformers, such as Stable Diffusion 3.

2. Preliminary

The Multi-Modal Diffusion Transformer (MM-DiT) employed in Stable Diffusion 3 uses three text encoders to create text embeddings in the following manner.



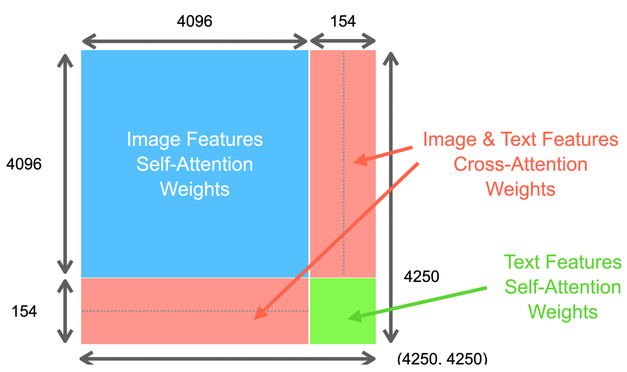
Text Embedding Generation



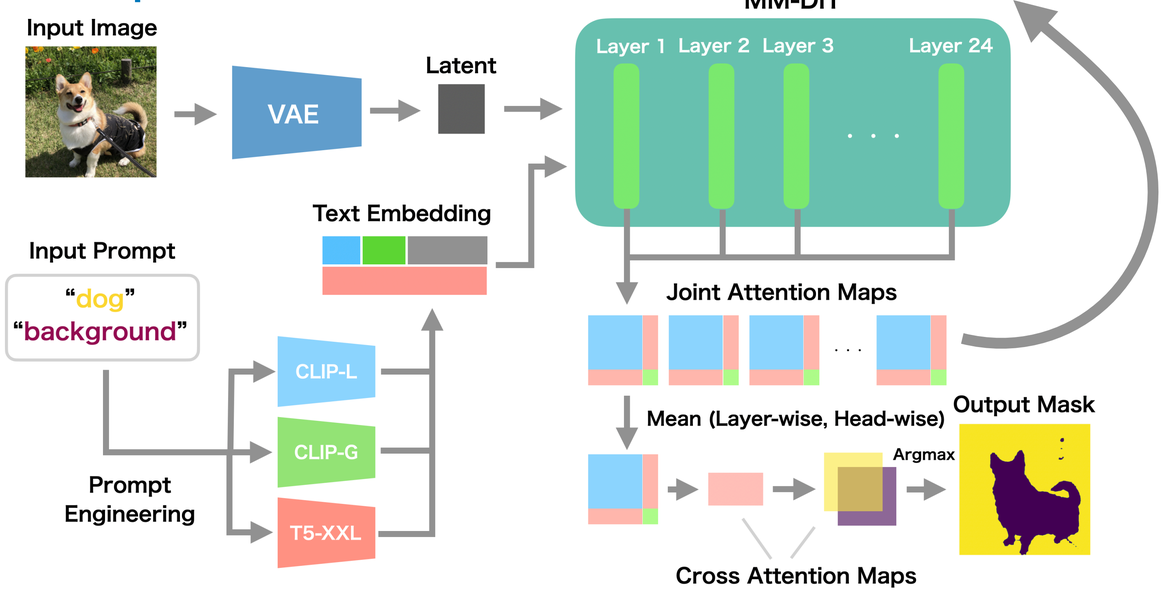
How to embed a object class in our method

3. Methodology

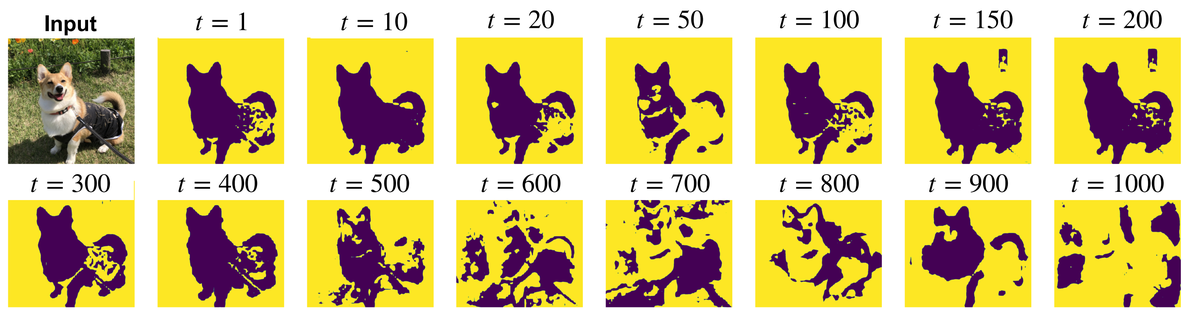
Joint Attention Map from MM-DiT's 1-step inference is extracted as shown on the right.



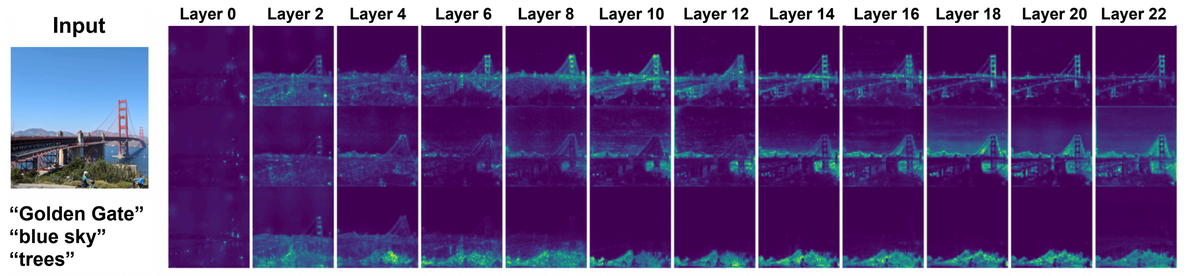
Our Pipeline



The regions vary significantly when altering the timestep or MM-DiT layers from which the Cross Attention Map is extracted.



The differences of segmentation results by timesteps



The differences of segmentation results by MM-DiT's layers

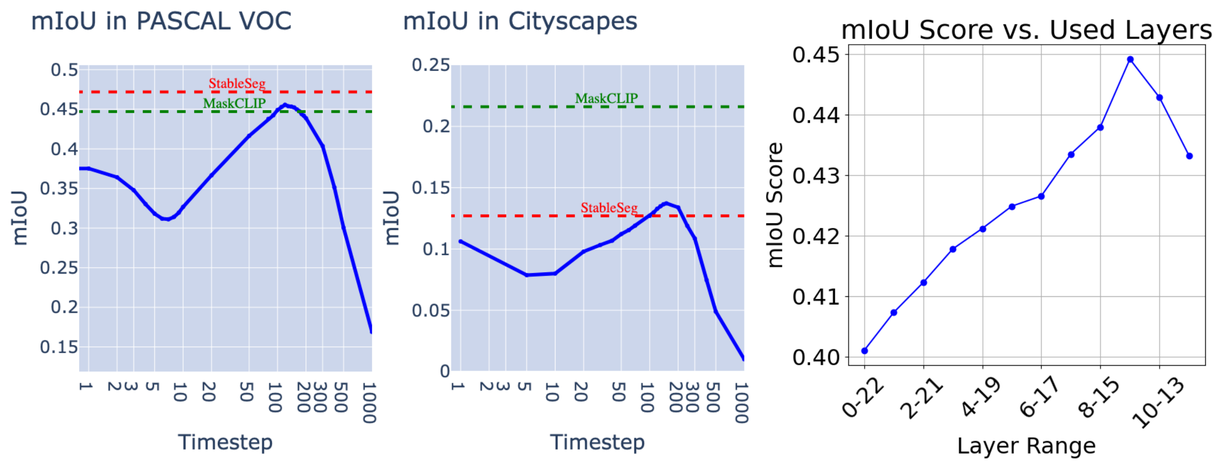
4. Experiments

Open-Vocabulary Segmentation

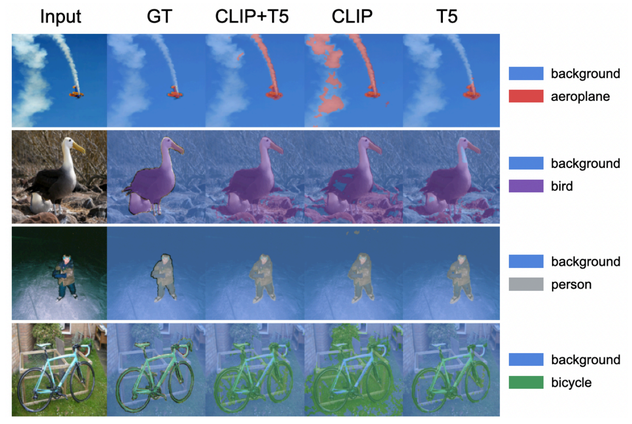
Our method enables segmentation of arbitrary objects by utilizing the Stable Diffusion 3 model.

Comparison on Hyperparameters

Evaluation on different timesteps and layers of MM-DiT and compared with existing training-free methods, MaskCLIP [1] and StableSeg [2].



Mehod	Pascal VOC	Cityscapes
Ours	0.452	0.137
MaskCLIP	0.447	0.216
StableSeg	0.472	0.127



Text Encoder Comparison

References:

[1] Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV. pp. 696–712. Springer Nature Switzerland, Cham (2022)

[2] Honbu, Y., Yanai, K.: Training-free region prediction with stable diffusion. In: ACM MM (2024)